Diffusion Models Beyond Mean Prediction

Mingtian Zhang

University College London

March 7, 2025

Diffusion Model

Forward Process: Let $q(x_0)$ be the data distribution. DDPM [1] assumes $q(x_t|x_0) = \mathcal{N}(\alpha_t x_0, \sigma_t^2 I)$ and Markovian forward process:

$$q(x_{0:T}) = q(x_0) \prod_{t=1}^{T} q(x_t|x_{t-1}),$$

where $\alpha_t^2 + \sigma_t^2 = 1$, we have $q(x_T) \to \mathcal{N}(0, I)$ when $T \to \infty$.

Reverse Process: Let $p(x_T) = \mathcal{N}(0, I)$, the Markovian generation process of DDPM:

1. Sample $x_T \sim p(x_T)$;
2. Sample $x_{t-1} \sim q(x_{t-1}|x_t)$ for $t = T, \cdots, 1$;

where $p_\theta(x_{t-1}|x_t) = \int q(x_{t-1}|x_0) q(x_0|x_t) \, \mathrm{d}x_0$ is the denoising model.

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. NeurIPS 2020

Understanding $q(x_0|x_t)$

- What do we know about $q(x_0|x_t)$?

Mean of $q(x_0|x_t)$

Let $q(x_t) = \int q(x_t|x_0)q(x_0)\,\mathrm{d}x_0$, where $q(x_t|x_0) = \mathcal{N}(\alpha_t x_0, \sigma_t^2 I)$, we have

Gaussian Identity

$$\nabla_{x_t} q(x_t|x_0) = \nabla_{x_t} \log q(x_t|x_0)q(x_t|x_0) = \frac{\alpha_t x_0 - x_t}{\sigma_t^2} q(x_t|x_0).$$

Tweedie's Lemma

$$\begin{aligned}
\nabla_{x_t} \log q(x_t) &= \frac{\nabla_{x_t} q(x_t)}{q(x_t)} = \frac{\int \nabla_{x_t} q(x_t|x_0)q(x_0)\,\mathrm{d}x_0}{q(x_t)} \\
&= \int \frac{\alpha_t x_0 - x_t}{\sigma_t^2} \frac{q(x_t|x_0)q(x_0)}{q(x_t)}\,\mathrm{d}x_0 \\
\Rightarrow \sigma_t^2 \nabla_{x_t} \log q(x_t) + x_t &= \int \alpha_t x_0 \frac{q(x_t|x_0)q(x_0)}{q(x_t)}\,\mathrm{d}x_0 = \alpha_t \langle x_0 \rangle_{q(x_0|x_t)} \\
\Rightarrow \langle x_0 \rangle_{q(x_0|x_t)} &= \frac{\sigma_t^2 \nabla_{x_t} \log q(x_t) + x_t}{\alpha_t}.
\end{aligned}$$

Understanding $q(x_0|x_t)$

- We know the mean of $q(x_0|x_t)$:

$$\langle x_0 \rangle_{q(x_0|x_t)} = \frac{\sigma_t^2 \nabla_{x_t} \log q(x_t) + x_t}{\alpha_t}.$$

- How to estimate $\nabla_{x_t} \log q(x_t)$?

Learning to Approximate $\nabla_{x_t} \log q(x_t)$

Denoising Score Identity For any convolution kernel $q(x_t|x_0)$, e.g., $q(x_t|x_0) = \mathcal{N}(\alpha_t x_0, \sigma_t^2)$, we have

$$\nabla_{x_t} \log q(x_t) = \frac{\int \nabla_{x_t} q(x_t|x_0) q(x_0)\, \mathrm{d}x_0}{q(x_t)} = \frac{\int \nabla_{x_t} \log q(x_t|x_0) q(x_t|x_0) q(x_0)\, \mathrm{d}x_0}{q(x_t)}$$
$$= \int \nabla_{x_t} \log q(x_t|x_0) q(x_0|x_t)\, \mathrm{d}x_0.$$

Denoising Score Matching

$$\mathrm{DSM}(\theta) = \frac{1}{2} \int q(x_t) ||\nabla_{x_t} \log q(x_t) - s_\theta(x_t)||_2^2\, \mathrm{d}x_t$$
$$\doteq \frac{1}{2} \iint q(x_t|x_0) q(x_0) ||\nabla_{x_t} \log q(x_t|x_0) - s_\theta(x_t)||_2^2\, \mathrm{d}x_t\, \mathrm{d}x_0$$
$$\doteq \frac{1}{2} \iint q(x_t|x_0) q(x_0) \left\| \frac{\alpha_t x_0 - x_t}{\sigma_t^2} - s_\theta(x_t) \right\|_2^2 \mathrm{d}x_t\, \mathrm{d}x_0.$$

Understanding $q(x_0|x_t)$

- We know the mean of $q(x_0|x_t)$:

$$\langle x_0 \rangle_{q(x_0|x_t)} = \frac{\sigma_t^2 \nabla_{x_t} \log q(x_t) + x_t}{\alpha_t}.$$

- We can learn to approximate $\nabla_{x_t} \log q(x_t)$ using DSM.
- Gaussian model approximation:

$$q(x_0|x_t) \approx p_\theta(x_0|x_t) \equiv \mathcal{N}(\mu_0(x_t), \Sigma_0(x_t)).$$

Heuristic Variance: DDPM

- Gaussian denoiser with learned mean and **fixed** variance:

$$p_\theta(x_{t-1}|x_t) = \int q(x_{t-1}|x_0)p_\theta(x_0|x_t)\,\mathrm{d}x_0 \equiv \mathcal{N}(\mu_{t-1}(x_t;\theta), \Sigma^2_{t-1}).$$

- Mean representation with **1st-order Tweedie's lemma**:

$$\mu_{t-1}(x_t;\theta) = \alpha_{t-1}\mu_0(x_t;\theta) = \alpha_{t-1}(\sigma_t^2 s_\theta(x_t,t) + x_t)/\alpha_t.$$

- Learn $s_\theta(x_t,t) \approx \nabla_{x_t} \log q(x_t)$ with Denoising Score Matching:

$$\mathrm{DSM}(\theta) = \frac{1}{2}\iint q(x_t|x_0)q(x_0)\left\|\frac{\alpha_t x_0 - x_t}{\sigma^2} - s_\theta(x_t)\right\|_2^2 \mathrm{d}x_t\,\mathrm{d}x_0.$$

- Two heuristic choice for the denoising variance:
  1. Set Σ_{t-1} to be the variance of $q(x_t|x_{t-1})$;
  2. Set Σ_{t-1} to be the variance of $q(x_{t-1}|x_t, x_0)$,

  both choices have similar results when $T \to \infty$.

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. NeurIPS 2020

Learning the Diagonal Covariance: I-DDPM

- Gaussian model with learned mean and fixed variance:

$$p_(x_{t-1}|x_t) = \int q(x_t|x_t)p_\theta(x_0|x_t) \equiv \mathcal{N}(\mu_{t-1}(x_t;\theta), \Sigma_{t-1}(x_t;\theta)).$$

- Learn $s_\theta(x_t, t) \approx \nabla_{x_t} \log q(x_t)$ for all t with DSM.
- Tweedie's Lemma for the mean estimation:

$$\mu_{t-1}(x_t;\theta) = \alpha_{t-1}\mu_0(x_t;\theta) = \alpha_{t-1}(\sigma_t^2 s_\theta(x_t, t) + x_t)/\alpha_t.$$

- Learn $\Sigma_{t-1}(x_t;\theta)$ fixed mean for all t with ELBO.
- Better FID and Likelihood with fewer NFEs.

[1] Alexander Quinn Nichol, and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. ICML 2021.

Estimating the Optimal Variance: A-DDPM

- Gaussian model with learned mean and fixed variance:

$$p_\theta(x_{t-1}|x_t) = \int q(x_t|x_0)p_\theta(x_0|x_t)\,\mathrm{d}x_0, p_\theta(x_0|x_t) \equiv \mathcal{N}(\mu_0(x_t;\theta), \tau_0^2 I).$$

- Learn $s_\theta(x_t, t) \approx \nabla_{x_t} \log q(x_t)$ for all t with DSM.
- Tweedie's Lemma for the mean estimation:

$$\mu_0(x_t;\theta) = (\sigma_t^2 s_\theta(x_t, t) + x_t)/\alpha_t.$$

- The optimal τ_0^* under ELBO has the form:

$$\tau_0^* = \left(\sigma_t^2 - \frac{\sigma_t^4}{d}\int \|s_\theta(x_t, t)\|_2^2\, q(x_t)\,\mathrm{d}x_t\right)/\alpha_t.$$

- I will come back.

[1] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models, ICLR 2022.

Why Can We Know More Beyond Mean Prediction?

Fisher Divergence:

$$\mathrm{DSM}(\theta) = \frac{1}{2}\int q(x_t)||\nabla_{x_t}\log q(x_t) - s_\theta(x_t)||_2^2\,\mathrm{d}x_t \equiv \mathrm{FD}(q(x_t)||p_\theta(x_t)),$$

where $p_\theta(x_t) = \int p(x_t|x_0)p_\theta(x_0)\,\mathrm{d}x_0$ and $\nabla_{x_t}\log p_\theta(x_t) = s_\theta(x_t)$.

Spread Fisher Divergence:

$$\mathrm{FD}(q(x_t)||p_{\theta^*}(x_t)) = 0 \Leftrightarrow q(x_t) = p_{\theta^*}(x_t) \Leftrightarrow q(x_0) = p_{\theta^*}(x_0),$$

for Gaussian $p(x_t|x_0)$, see paper [1] for a proof.

Implication: *Given $q(x_t|x_0)$, $\nabla_{x_t}\log q(x_t)$ fully characterizes $q(x_0)$. We then have all the information of $q(x_0|x_t)$, not just mean.*

[1] Mingtian Zhang, Peter Hayes, Thomas Bird, Raza Habib, and David Barber. Spread Divergence. ICML 2020.

Moment Matching Approximation

High-dimensional 2nd-order Tweedie's lemma [1]

Let $q(x_t) = \int q(x_t|x_0)q(x_0)\,\mathrm{d}x_0$, $q(x_t|x_0) = \mathcal{N}(\alpha_t x_0, \sigma_t^2 I)$, we have

$$\nabla_{x_t}^2 \log q(x_t) = \nabla_{x_t} \frac{\nabla_{x_t} q(x_t)}{q(x_t)} = \nabla_{x_t} \frac{\int \nabla_{x_t} q(x_t|x_0)q(x_0)\,\mathrm{d}x_0}{q(x_t)}$$

$$\Longrightarrow \frac{\sigma_t^4 \alpha_t {}_{x_t}^2 \log q(x_t) + \sigma_t^2 I}{\alpha_t^2} = \langle x_0^2 \rangle_{q(x_0|x_t)} - \langle x_0 \rangle_{q(x_0|x_t)}^2 \equiv \Sigma_0(x_t).$$

Score-based moment matching approximation [2]

Assume $p_\theta(x_0|x_t) = \mathcal{N}(\mu_0(x_t;\theta), \Sigma_0(x_t;\theta))$, where

$$\mu_0(x_t;\theta) = (\sigma_t^2 s_\theta(x_t) + x_t)/\alpha_t,$$
$$\Sigma_0(x_t;\theta) = (\sigma_t^4 \nabla_{x_t} s_\theta(x_t) + \sigma_t^2 I)/\alpha_t^2.$$

[1] Bradley Efron. Tweedie's formula and selection bias. JASA 2011.

[2] Mingtian Zhang, Alex Hawkins-Hooker, Brooks Paige, and David Barber. Moment matching denoising gibbs sampling, NeurIPS 2024.

Connections

A-DDPM: $p_\theta(x_0|x_t) = \mathcal{N}(\mu_0(x_t;\theta), \tau_0^2 I)$

$$\tau^{*2} = \arg\min_{\sigma_q} \mathrm{KL}(q(x_t|x_0)q(x_0)\|p_\theta(x_0|x_t))q(x_t)) \tag{1}$$

$$= \frac{1}{d}\left\langle \mathrm{Tr}\left(\mathrm{Cov}_{q(x_0|x_t)}[x_0]\right)\right\rangle_{q(x_t)}, \tag{2}$$

SN-DDPM/NPR-DDPM: $p_\theta(x_0|x_t) = \mathcal{N}(\mu_0(x_t;\theta), \Sigma_0(x_t;\theta))$:

$$\mathrm{Cov}_{q(x_0|x_t)}[x_0] = \frac{\beta_t^2}{\alpha_t}\left(\frac{1}{\beta_t}\mathbb{E}_{q(x_0|x_t)}[\epsilon\epsilon^T] - \nabla_{x_t}\log q(x_t)\nabla_{x_t}\log q(x_t)^T\right),$$

where $\epsilon = (x_t - \alpha_t x_0)/\sigma_t$. Learn $\Sigma_0(x_t;\theta)$ to approximate $\mathrm{Cov}_{q(x_0|x_t)}[x_0]$.

[1] Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang, Estimating the optimal covariance with imperfect mean in diffusion probabilistic models, ICML 2022

Diagonal Approximation and Amortization

Hutchinson trick for diagonal Hessian approximation

$$\begin{aligned}\operatorname{diag}(\nabla^2_{x_t} \log q(x_t)) &\approx 1/M \sum_{m=1}^{M} v_m \odot \nabla^2_{x_t} \log q(x_t) v_m \\ &\approx 1/M \sum_{m=1}^{M} v_m \odot \nabla_{x_t}(s_\theta(x_t)^T v_m),\end{aligned}$$

where $v \sim p(v)$ is a random vector taking values in $\{+1, -1\}$.

Neural network amortization $h_\phi(x_t)$ with biased regression

$$L(\phi) = \int p(x_t) \left\| h_\phi(x_t) - \int v \odot \nabla_{x_t}(s_\theta(x_t)^T v) p(v)\, \mathrm{d}v \right\|_2^2 \mathrm{d}x_t.$$

Unbiased regression: Optimal Covariance Matching (OCM)

$$L(\phi) = \int p(x_t) p(v) \left\| h_\phi(x_t) - v \odot \nabla_{x_t}(s_\theta(x_t)^T v) \right\|_2^2 \mathrm{d}v\, \mathrm{d}x_t.$$

[1] Zijing Ou*, Mingtian Zhang*, Andi Zhang, Tim Z. Xiao, Yingzhen Li, and David Barber. Improving Probabilistic Diffusion Models With Optimal Covariance Matching, ICLR 2025.

Overview

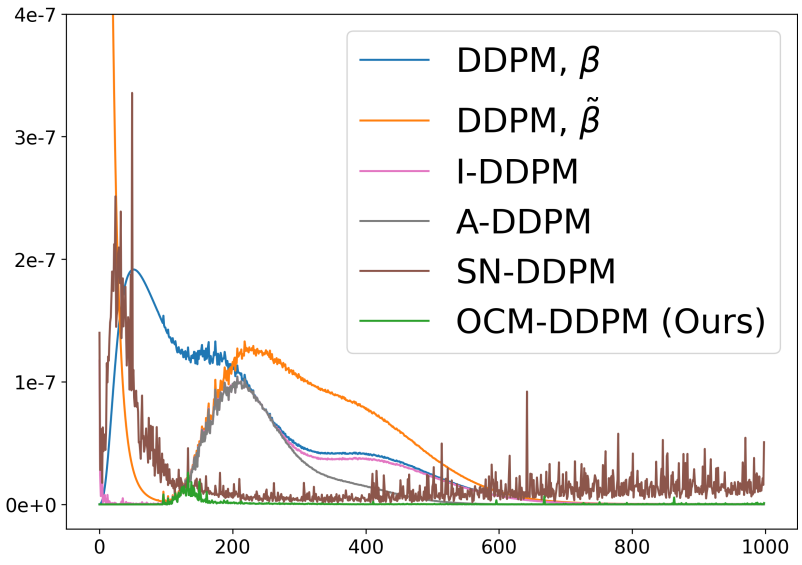
Table: Overview of different covariance estimation methods. We include the intuition of the methods and how many additional network passes are required for estimating the covariance.

Modeling Capability ↓

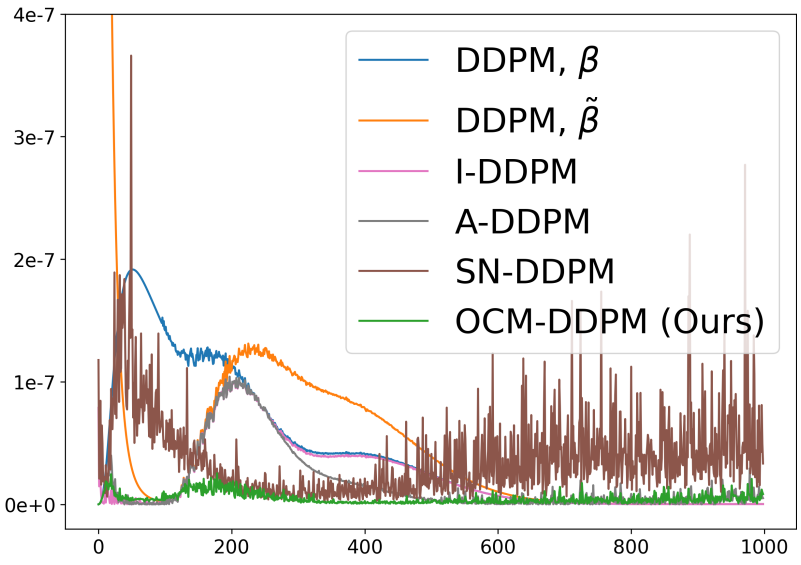
Covariance Type	**+#Passes**	**Intuition**
x_t-independent Isotropic: β-DDPM	0	Cov. of $q(x_t\|x_{t-1})$
x_t-independent Isotropic: $\tilde{\beta}$-DDPM	0	Cov. of $q(x_t\|x_{t-1}, x_0)$
x_t-independent Isotropic Estimation: A-DDPM	0	Estimate from data
x_t-dependent Diagonal VLB: I-DDPM	1	Learn from data
x_t-dependent Diagonal: NS-DDPM	1	Learn from data
x_t-dependent Diagonal: OCM-DDPM	1	Learn from score
x_t-dependent Diagonal Estimation: Hutchinson	$2M$	Estimate from score
x_t-dependent Diagonal Analytic: Exact	D	Calculate from score
x_t-dependent Full Analytic: Exact	D	Calculate from score

Covariance Estimation Accuracy

Data: Two-dimensional mixture of 9 Gaussians (MoG) with means located at $\{-3, 0, 3\} \otimes \{-3, 0, 3\}$ and $\sigma = 0.1$.



(a) Cov Error with True Score



(b) Cov Error with Learned Score

Applications in Diffusion Acceleration

Given **OCM** approximation

$$p_\theta(x_0|x_t) = \mathcal{N}\left(\mu_0(x_t;\theta), \Sigma_0(x_t;\theta)\right),$$

where

$$\mu_0(x_t;\theta) = (\sigma_t^2 s_\theta(x_t) + x_t)/\alpha_t, \quad \Sigma_0(x_t;\phi) = (\sigma_t^4 h_\phi(x_t) + \sigma_t^2 I)/\alpha_t^2.$$

Skip-Step DDPM

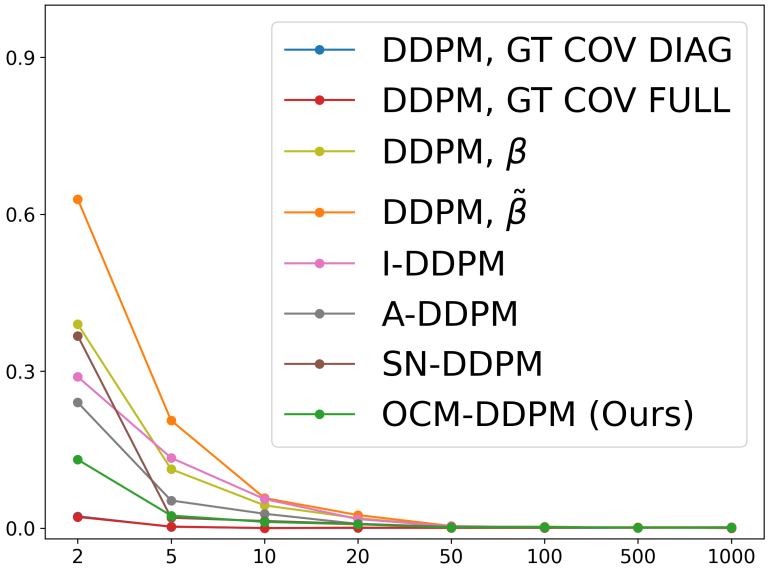
$$p_\theta(x_{t-\Delta}|x_t) = \int q(x_{t-\Delta}|x_0) p_\theta(x_0|x_t)\,\mathrm{d}x_0,$$

Skip-Step DDIM

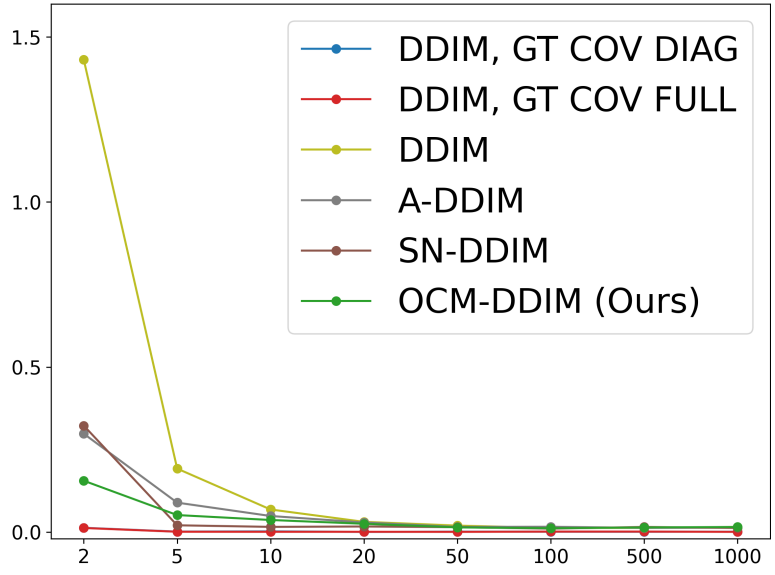
$$p(x_{t-\Delta}|x_t) = \int q(x_{t-\Delta}|x_t, x_0) p_\theta(x_0|x_t) dx_0,$$

Covariance Flexibility I

Data: Two-dimensional mixture of 9 Gaussians (MoG) with means located at $\{-3, 0, 3\} \otimes \{-3, 0, 3\}$ and $\sigma = 0.1$.



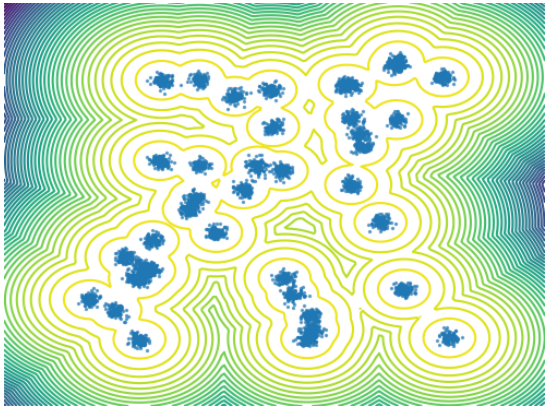
(a) DDPM MMD v.s. Steps



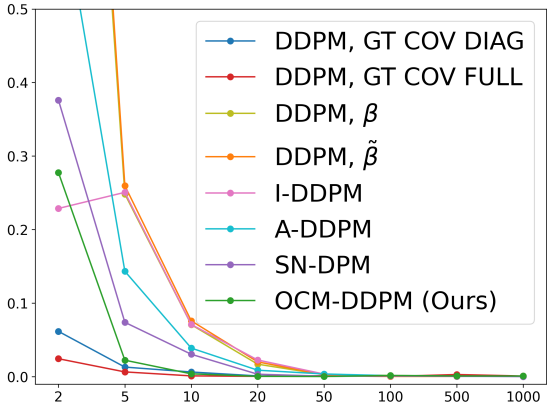
(b) DDIM MMD v.s. Steps

Figure: In this example, performance is dominated by the diagonal convergence.

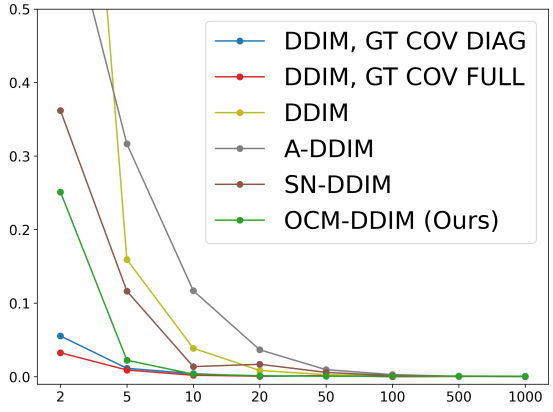
Covariance Flexibility II



(a) Training Data



(b) DDPM



(c) DDIM

Figure: In this example, full covariance is better than diagonal covariance.

Image Experiment

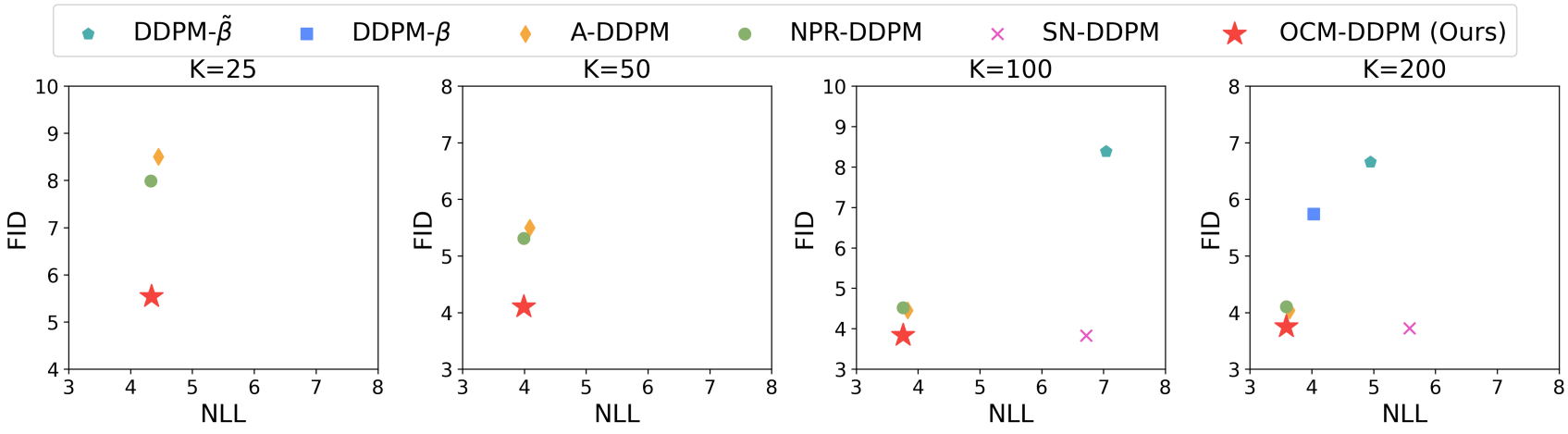


Figure: The results of FID v.s. NLL for different methods with varying numbers of sampling steps on CIFAR10 (CS). Our method consistently achieves the best trade-off between FID and NLL.

Image Experiment

METHOD	CIFAR10	CELEBA 64X64	LSUN BEDROOM	IMAGENET 256X256
DDPM	90 (6.12)	> 200	130 (6.06)	21 (5.89)
DDIM	30 (5.85)	> 100	BEST FID > 6	11 (5.58)
IMPROVED DDPM	45 (5.96)	MISSING MODEL	**90** (6.02)	22 (6.08)
ANALYTIC-DPM	25 (5.81)	55 (5.98)	100 (6.05)	MISSING MODEL
NPR-DPM	1.002×23 (5.76)	1.013×50 (6.04)	$1.021\times$**90** (6.01)	MISSING MODEL
SN-DPM	1.005×17 (5.81)	1.019×22 (5.96)	1.114×92 (6.02)	MISSING MODEL
OCM-DPM (OURS)	$1.003\times$**16** (5.83)	$1.015\times$**21** (5.94)	$1.112\times$**90** (6.04)	$1.007\times$**10** (5.33)

Figure: The least number of timesteps $\downarrow$ required to achieve an FID around 6 (along with the corresponding FID).

Image Experiment

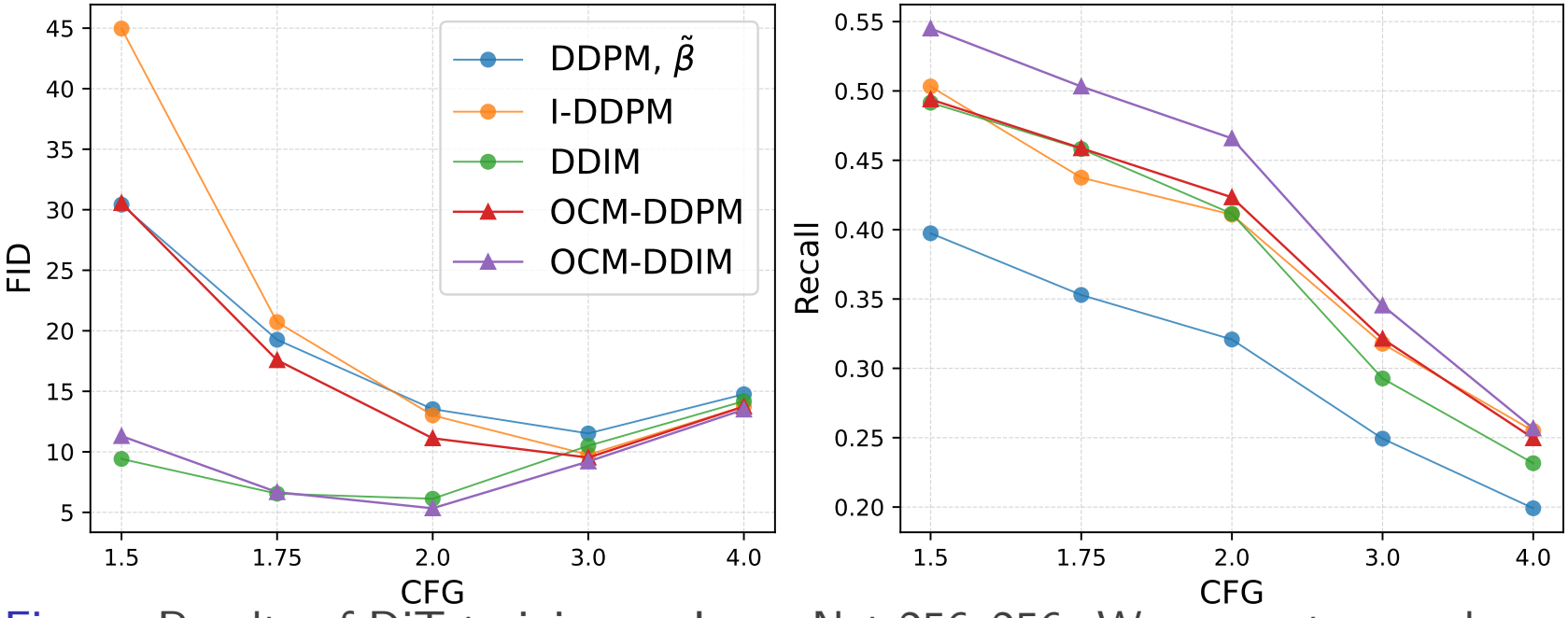


Figure: Results of DiT training on ImageNet 256x256. We generate samples using 10 timesteps with varying CFG coefficients.

Discussions

- A better $q(x_{t-1}|x_t)$ can improve generation quality and likelihood with fewer steps.
- The Gaussian assumption of the denoising posterior remains a constraint. A recent paper [1] directly learns a neural sampler to approximate $q(x_{t-1}|x_t)$.
- Is continuous diffusion acceleration still an open research problem? One-step generation methods have achieved competitive or even better quality than multi-step diffusion.
- Check out our work [2] on one-step generative models.

[1] Valentin De Bortoli, Alexandre Galashov, J. Swaroop Guntupalli, Guangyao Zhou, Kevin Murphy, Arthur Gretton, and Arnaud Doucet. Distributional Diffusion Models with Scoring Rules, https://arxiv.org/abs/2502.02483.

[2] Mingtian Zhang*, Jiajun He*, Wenlin Chen*, Zijing Ou, José Miguel Hernández-Lobato, Bernhard Schölkopf, David Barber, Towards Training One-Step Diffusion Models Without Distillation, https://mingtian.ai/pdf/OneStepModel.pdf.